THEORY-BASED EVALUATION

Contents

Based on material produced for DG Regional Policy by Frans L. Leeuw

1. Introduction

Over the last forty years a number of evaluation experts (Suchman, 1967; Chen and Rossi, 1980; Weiss, 1995; Pawson and Tilley, 1997; Rogers et al., 2008; Donaldson, 2007) have contributed to -the development of what can be called a theory-oriented evaluation approach - called theory-based evaluation, theory-driven evaluation or programme theory evaluation. For the purposes of EVALSED, the term theory-based evaluation (TBE) is used to reflect these approaches.

The objective of this guidance is to provide users of EVALSED with some general ideas of what TBE is, what questions it can answer under which circumstances and how the approach can be applied, using various evaluations methods.

Several approaches have been developed within TBE over the years. However, these approaches have not been applied often within the socio-economic development programmes financed under EU Cohesion policy. Therefore, the present guidance provides examples of how TBE has been used in other intervention fields. When good practice examples are available in the field of the EU Cohesion policy, the guidance material will be updated.

Some of the data collection techniques relevant for TBE, e.g., focus groups, workshops, case studies, expert judgements, are explained in section 2. More information on data collection is included on EVALSED under the Sourcebook: Method and Techniques in the Section: Collecting Information:

http://ec.europa.eu/regional_policy/sources/docgener/evaluation/evalsed/sourcebooks/method_techniques/collecting_information/index_en.htm

Users of the guidance are encouraged to refer to the examples and references given before applying the approaches for the first time. The approaches build upon a wealth of experience and literature that is not fully reviewed here. Therefore, it would be advisable, especially for those who wish to explore TBE in more detail, to refer to additional expertise and other specialised literature.

Definition of Theory Based Evaluation

Theory-based evaluation is an approach in which attention is paid to *theories* of policy makers, programme managers or other stakeholders, i.e., collections of assumptions, and hypotheses - empirically testable - that are logically linked together.

These theories can express an intervention logic of a policy: policy actions, by allocating (spending) certain financial resources (the inputs) aim to produce planned outputs through which intended outcomes in terms of people's well-being and progress[1] are expected to be achieved. The actual outcomes will depend both on policy effectiveness and on other factors affecting outcomes, including the context. An essential element of policy effectiveness is the mechanisms that make the intervention work. Mechanisms are not the input-output-outcome chain, the logic model or statistical equations. They concern amongst others beliefs, desires, cognitions and other decision-making processes that influence behavioural choices and

[1] In this guidance, the meaning of the term 'outcome' is the same as 'result' used in the European Commission publications related to the EU Cohesion policy in the upcoming programming period 2014-2020. It means a specific dimension of well-being and progress for people that motivates policy action or happens as a consequence of policy action, i.e. what is intended to be changed or what has changed, with the contribution of the interventions designed.

actions. Theory based evaluation explores the mechanisms which policy makers believe make the policy effective and compares these with research based evidence.

Theory-based evaluation focuses on this intervention theory; it aims to find and articulate this theory, to test it and to improve it, if necessary.

Theory-based evaluation has at its core two vital components. The first is conceptual, the second empirical. Conceptually, theory-based evaluations articulate a policy or programme theory. Empirically, theory-based evaluations seek to test this theory, to investigate whether, why or how policies or programmes cause intended or observed outcomes.

Testing the theories can be done on the basis of existing or new data, both quantitative (experimental and non-experimental) and qualitative. TBE does not apply a hierarchy of research designs and methods; it does not favour any over any others, as long as they are rigorously applied. Their choice depends on the evaluation design and they should be selected if they are appropriate to answer the evaluation questions.

2. How to Articulate and Test Policy or Programme Theories? Main Approaches Used

Theories underlying a policy or programme are often not directly visible or knowable to evaluators. They are often not explicitly expressed in official documents. Evaluators have to search for these theories – if they have not been concisely articulated - and explain them in a testable way. Then they have to test them. Below are briefly presented some approaches to how to do this. The list is not exhaustive.

2.1 Realist Evaluation

The term realist evaluation was coined by Ray Pawson and Nick Tilley in their book with the same title (1997). This methodological approach stresses the importance of CMO (Context, Mechanism, Outcomes) configurations basic to policies and programmes.

Let us take an example of a socio-economic development programme. Such a programme attempts to solve a problem – to create some kind of socio-economic change. The programme works by enabling stakeholders to make choices. But choice-making is always constrained by stakeholders' previous experiences, beliefs and attitudes, opportunities and access to resources.

Making and sustaining different choices requires a change in stakeholders' reasoning (for example, values, beliefs, attitudes or the logic they apply to a particular situation) and the resources (e.g. information, skills, material resources, financial support) available to them. This combination of reasoning and resources is what enables the programme to work and is known as a programme ‘mechanism’. The programme works in different ways and sometimes for different people (that is, the programme can trigger different change mechanisms for different stakeholders).

The contexts in which the programme operates make a difference to the outcomes it achieves. Programme contexts include features such as social, economic and political structures, organisational context, programme stakeholders, programme staffing, geographical and historical context and so on. Some factors in the context may enable particular mechanisms to be triggered. Other aspects of the context may prevent particular mechanisms from being triggered. There is always an interaction between context and mechanism and that interaction is what creates the programme's outcomes: Context + Mechanism = Outcome.

Because programmes work differently in different contexts and through different change mechanisms, they cannot simply be replicated from one context to another and automatically achieve the same outcomes. Knowledge about 'what works for whom, in what contexts, and how' are, however, portable.

Therefore, one of the tasks of evaluation is to learn more about 'what works for whom', 'in which contexts particular programmes do and don't work', and 'what mechanisms are triggered by what programmes in what contexts'.

A realist approach assumes that programmes are theories incarnate. That is, whenever a programme is implemented, it is testing a theory about what might cause change, even though that theory may not be explicit. One of the tasks of a realist evaluation is, therefore, to make the theories within a programme explicit, by developing clear hypotheses about how, and for whom, programmes might 'work'. The implementation of the programme, and the evaluation of it, then tests those hypotheses. This means collecting data, quantitative and qualitative, not just about programme outcomes, or the processes of programme implementation, but about the specific aspects of programme context that might impact on programme outcomes, and about the specific mechanisms that might create change.

Pawson and Tilley also argue that a realist approach has particular implications for the design of an evaluation and the roles of those benefiting from a programme. For example, rather than comparing changes for participants who have benefitted from a programme with a group of people who have not (as is done in random control or non-experimental designs), a realist evaluation compares mechanisms and outcomes within and between programmes. It may ask, for example, whether a programme works differently in different localities (and if so, how and why); or for different population groups (for example, men and women, or groups with differing socio-economic status). They argue that different stakeholders will have different information and understandings about how programmes are supposed to work and whether they in fact do. Data collection processes (interviews, focus groups, collection of administrative data, questionnaires and so on) should be constructed to collect the particular information that those stakeholder groups will have, and thereby, to refute or refine theories about how and for whom the programme works.

Pawson and Sridharan (2010) present the following methodological steps for a realistic evaluation of a programme:

- *Eliciting and surfacing the underlying programme theories*

The first point is that although programme theories 'are easily spotted', these theories are best elicited from their procreators and this may involve either:

- Reading and closely analysing programme documentation, guidance, regulations, etc., on how the programme will achieve its ends, or
- Interviews with programme architects, managers or practitioners on how their intervention will generate the desired change.

Programme theories normally flow quite readily from these interviews, though some related difficulties should be noted. The first, located at the political level, is a tendency to ambiguity in policy discourse. The second problem, located nearer the programme practice, occurs when the core theory is either seemingly so obvious or buried tacitly in the minds of the programme makers that it can fail to surface in the interview. In these situations, persuasion is sometimes needed to encourage practitioners to spell out how their actions worm their way into participants' choices (Pawson and Tilley 1997).

- *Mapping and selecting the theories to put to research*

Having found a means to elicit the programme theories at work, the next stage is to begin to codify or map them. An array of techniques is available for this task, known variously as concept mapping, logic modelling, system mapping, problem and solution trees, scenario building, configuration mapping, and so on. All try to render the process through which the programme achieves its ends, usually in diagrammatic form. These maps may identify the various causes of the problem, the administrative steps to take, the unfolding sequence of programme activities and inputs, the successive shifts in dispositions of participants, the progressive targeting of programme recipients, the sequence of intervention outputs and outcomes.

- *Formalising the theories to put to test*

After eliciting, mapping and selecting programme theories, the time comes to formalise them. They need to be transformed into a propositional form, as hypotheses suitable for empirical research. Programme theories come to life as insights, brain waves, bright ideas, and informed guesses. Sometimes, they turn out to be wishful thinking and pipe dreams. What evaluation research requires, by contrast, are testable propositions.

- *Data collection and analysis*

Data collection and analysis follows an empirical research (qualitative and quantitative, including experimental and non-experimental techniques) in order to understand, test and refine policy or programme theories (regarding CMO).

Realist evaluators refer to ´interrogating the policy or programme theories'; they also stress that whilst it was appropriate to follow custom and refer to this stage as 'theory testing', these approaches prefer the term 'theory refinement'. The objective is not to accept or reject policy/programme theories. The mission is to improve them.

2.2 Theory of Change

Carol Weiss (1995) popularized the term 'theory of change'. She hypothesized that a key reason complex policies or programmes are so difficult to evaluate is that the assumptions that inspire them are poorly articulated. She argued that stakeholders of complex initiatives are typically unclear about how the change process will unfold and therefore pay little attention to the early and mid-term changes that need to happen in order for a longer term goal to be reached. The lack of clarity about the 'mini-steps' that must be taken to reach a long term outcome not only makes the task of evaluating a complex initiative challenging, but reduces the likelihood that all important factors related to the long term goal will be examined. Weiss defined theory of change as a way to describe the set of assumptions that explain both the mini-steps that lead to the long term goal and the connections between policy or programme activities and outcomes that occur at each step of the way. She challenged designers of complex initiatives, such as EU programmes, to be specific about the theories of change guiding their work and suggested that doing so would improve their policies and would strengthen their ability to claim credit for outcomes that were predicted in their theory.

The following steps elicit the theory of change underlying a planned programme. A pre-condition is that the evaluator works collaboratively with a wide range of stakeholders.

Step 1: The focus is on the long-term vision of a programme and is likely to relate to a timescale that lies beyond its timeframe. Its aim should be closely linked to the existence of a local, regional or national problem. For example, a smoking cessation programme might have a long-term vision of eradicating inequalities in smoking prevalence by 2020.

Step 2: Having agreed the ultimate aim of the programme, stakeholders are encouraged to consider the necessary outcomes that will be required by the end of the programme if such an aim is to be met in the longer term. Within a programme they might, for instance, anticipate a decrease in gap between the most and least deprived areas.

Steps 3 and 4: Stakeholders are asked to articulate the types of outputs and short-term outcomes that will help them achieve the specified targets. These might include reductions in differential access to acceptable smoking cessation programmes. At this stage those involved with the programme consider the most appropriate activities or interventions required to bring about the required change. Different strategies of engagement might be used to target pregnant women, middle-aged men and young adolescents, for example.

Step 5: Finally, stakeholders consider the resources that can realistically be brought to bear on the planned interventions. These will include staff and organisational capacity, the existence of supportive networks and facilities as well as financial capability.

Following a collective and iterative process the resulting programme theory must fulfill certain criteria: it must be plausible, doable and testable. It needs to be articulated in such a way that it can be open to evaluation; this is only possible where there is a high degree of specificity concerning the desired outcomes. Only then, the theory of change that is elicited should be interrogated to ensure that the underlying logic is one that is acceptable to stakeholders either because of the existing evidence base or because it seems likely to be true in a normative sense. The evaluator then takes the programme map generated through this process and, using various data (qualitative and quantitative) collection techniques as relevant, monitors and analyses the unfolding of the programme in practice and integrates the findings.

2.3 Contribution Analysis

'Contribution analysis' is a performance measurement approach developed within the Office of the Canadian Auditor General in the 1990s and it aims to establish the contribution a programme makes to desired outcomes.

In practice, many evaluations identify whether or not an outcome has been achieved and, if it was, assume the programme can take credit for this. However, reporting on outcomes and proving attribution are two different things. Attribution involves drawing causal links and explanatory conclusions between observed changes and specific interventions. Determining whether an outcome was caused by a programme or other external factors is difficult and expensive. However, demonstrating the contribution of a programme to outcomes is crucial if the value of the programme is to be demonstrated and to enable decisions to be made about its future direction (Mayne 2001).

Rather than attempt to definitively causally link a programme to desired outcomes, contribution analysis seeks to provide plausible evidence that can reduce uncertainty regarding the difference a programme is making to observed outcomes (Mayne 2001).

The following description is based on Mayne (2010) where seven methodological steps are described that form a contribution analysis.

Step 1: Set out the cause-effect issue to be addressed

First, it is necessary to articulate cause and effect. The following questions should be posed:

- Would the expected intervention make a difference to the problem?
- What aspects of the intervention or the context would lead to a contribution being made?
- What would provide evidence that the intervention made a noticeable contribution?
- Is the expected contribution plausible given the nature of the intervention and the problem being addressed? If not, the value of further analysis needs to be reassessed.

Step 2: Develop the theory of change

Developing a prior or initial theory of change for the intervention is the second key step. Contribution analysis needs reasonably straightforward, not overly detailed outcomes chains, especially at the outset. Refinements may be needed to further explore some aspects of the theory of change, but can be added later.

Step 3: Assess the resulting contribution story

At this point, it is useful to critically review the contribution story resulting from the developed theory of change, i.e.:

- To assess the logic of the links and test the plausibility of the assumptions in the theory of change: Are there any significant gaps in the theory? Can they be filled by refining the theory of change? If not, is it worth continuing?
- To identify where evidence is needed to strengthen the contribution story: Which links have little evidence? Which external factors are not well understood?
- To determine how much the theory of change is contested: Is it widely agreed? Are specific aspects contested? Are there several theories of change at play?

Step 4: Gather existing evidence on the theory of change

Before gathering new data and information, it is useful and cost-effective to look at the relevant existing data and information there is about the theory of change. The purpose is to provide empirical evidence for the contribution story: evidence on activities implemented, on observed outcomes, on assumptions being realised and on relevant external factors.

At this point in the analysis, a theory of change for the intervention has been developed and the available evidence supporting the theory of change collected. The theory of change has to some extent been tested. The significant external factors have also been identified and available evidence gathered for them.

Step 5: Re-assess the contribution story and challenges to it

The theory of change can be critically assessed in light of the existing evidence:

- Which links in the theory of change are strong (strong logic, good evidence available supporting the assumptions, low risk and wide acceptance) and which are weak?
- How credible is the story overall? Does the pattern of outcomes and links between them validate the contribution chain?
- Do stakeholders agree with the contribution story developed?
- Is it likely that any of the external significant factors have had a noteworthy influence on the outcomes observed?
- What are the main weaknesses in the story? Where would additional data or information be useful?

Step 6: Seek out additional empirical evidence

This is the step where the primary data gathering for the evaluation begins, informed by the previous steps e.g., step 5 has identified where additional evidence is needed.

- Evidence is gathered to strengthen the contribution story, using appropriate data gathering techniques, such as surveys and reviewing and analysing administrative data. There may be evidence on outcomes occurring, on the validity of the assumptions and risks in the theory of change and on significant external factors.
- There may possibilities to use quantitative techniques (experimental and non-experimental designs) involving comparison groups that could be used to explore elements of the theory of change.
- From a theory-based perspective, several frequently used data gathering techniques can be strengthened:

– *Key informant interviews* can both test the theory of change developed and elicit alternative theories of change the key informants might have, as well as discuss other influencing factors. Interviewees should be asked what on evidence they base their views.

– *Focus groups* and *workshops* can explore a theory of change since there will be discussion about how different people see the intervention working. Alternative theories of change may emerge and other influencing factors can be identified. They can be used to develop a theory of change and as a way to identify evidence on the extent to which the theory of change has been realised in practice.

– *Case studies* can be used in the same way. Case studies are powerful as a data gathering tool to help confirm or refute a theory of change, or the micro steps in a theory of change, showing that the theory is indeed plausible and not just based on unsupported beliefs.

Step 7: Revise and strengthen the contribution story

Now, the newly collected empirical evidence should be used to build a more credible contribution story with strengthened conclusions on the causal links in the theory of change. Contribution analysis works best as an iterative process. At this point, the analysis may return to Step 5 and reassess the strengths and weaknesses of the contribution story and decide if further analysis is useful or possible.

2.4 Policy Scientific Approach

The 'policy scientific approach' covers the following six steps (Leeuw, 2003):

Step 1: Identify behavioral mechanisms expected to solve the problem

Searching in formal and informal documents and in interview transcripts can elicit statements that indicate why it is believed necessary to solve the policy problem and what the goals are of the policy or programme under review. These statements point to mechanisms; these can be considered the 'engines' that drive the policies or programmes and are believed to make them effective.

Step 2: Statements that have the following form are especially relevant for detecting these mechanisms:

- 'It is evident that *x* . . . will work'
- 'In our opinion, the best way to go about this problem is to . . .'
- 'The only way to solve this problem is to . . .'
- 'Our institution's *x* years of experience tells us that . . .';

Step 3: Compile a survey of these statements and link the mechanisms with the goals of the programme or policy under review

Step 4: Reformulate these statements in conditional 'if–then' propositions or propositions of a similar structure ('the more x, the less y').

Step 5: Search for 'warrants' to identify missing links in or between different propositions through argumentation analysis.

Argumentation analysis is a standard tool in logic and philosophy. It describes a model for analysing chains of arguments and it helps to reconstruct and fill in argumentations. A central concept is the 'warrant', the 'because' part of an argument: it says that B follows from A because of a (generally) accepted principle. For example, 'the organisation's performance will not improve next year' follows from 'the performance of this organisation has not improved over the last 5 years,' because of the principle, 'past performance is the best predictor of future performance.' The 'because' part of such an argument often is not made explicit. Consequently, these warrants must be inferred by the person performing the analysis

Step 6: Reformulate these 'warrants' in terms of conditional 'if–then' (or similar) propositions and draw a chart of the (mostly causal) links.

Step 7: Evaluate the validity of the propositions by looking into:

- the logical consistency of the set of propositions;
- their empirical content, that is, the extent to which the theory and, in particular, the assumed impact of the behavioral mechanisms correspond with the state of the art within the social/behavioral/economic sciences on these mechanisms.

Evaluating the reconstructed programme theory can be done in different ways. One is to confront (or juxtapose) different theories (like Carvalho & White, 2004, with regard to social funds). Another is to empirically test the programme theory by making use of primary or secondary data (triangulation), both qualitative and quantitative. A third possibility is to organise an iterative process of continuous refinement using stakeholder feedback and multiple data collection techniques and sources (in the realist tradition), while a fourth approach is to make use of already published reviews and synthesis studies. These can play a pivotal role in marshalling existing evidence to deepen the power and validity of a TBE, to contribute to future knowledge building and to meet the information needs of stakeholders. Visualisation or mapping software can help in this task.

There are several techniques for data collection and analysis, for example:

- Systematic reviews are syntheses of primary studies that, from an initial explicit statement of objectives, follow a transparent, systematic and replicable methodology of literature search, inclusion and exclusion of studies according to clear criteria, and extracting and synthesizing of information from the resulting body of knowledge.

- Meta-analyses quantitatively synthesize 'scores' for the impact of a similar set of interventions from a range of studies across different environments.

- Realist syntheses collect earlier research findings by placing the policy instrument or intervention that is evaluated in the context of other similar instruments and describe the intervention in terms of its context, social and behavioral mechanisms (what makes the intervention work) and outcomes.

2.5 Strategic Assessment Approach

Central in the Strategic Assessment Approach are four major stages: (1) group formation; (2) assumption surfacing; (3) dialectical debate; and (4) synthesis (Leeuw, 2003; Mason and Mitrof, 1980).

Stage 1 - Group Formation: The aim is to structure groups so that the productive operation of the later stages of the methodology is facilitated. A wide cross-section of individuals with an interest in the relevant policy question should be involved. They are divided into groups, care being taken to maximise convergence of viewpoints within groups and to maximise divergence of perspectives between groups.

Stage 2 – Assumption Surfacing: The different groups separately unearth the most significant assumptions that underpin their preferred policies or programmes. Two techniques assume importance in assisting this process.

The first, stakeholder analysis, asks each group to identify the key individuals or groups upon whom the success or failure of their preferred strategy would depend. This involves asking questions such as: Who is affected by the strategy? Who has an interest in it? Who can affect its adoption, execution, or implementation? And who cares about it? For the stakeholders identified, each group then lists what assumptions it is making about each of them in believing that its preferred strategy will succeed.

The second technique is assumption rating. Initially one should find and list the assumptions. This involves searching for statements about symptoms of the problem (that have to be solved through a policy or programme, distinguishing them from statements about causes of the problem). For each of the listed assumptions, each group asks itself two questions: (1) How important is this assumption in terms of its influence on the success or failure of the strategy? And (2) how certain are we that the assumption is justified? Here, in fact, the evaluation of the listed assumptions takes place, usually by using research reviews and similar documents. The results are recorded on a chart. Each group then is able to identify a number of key assumptions upon which the success of its strategy rests.

Stage 3 - Dialectical debate: The groups are brought back together and each group makes the best possible case to the others for its preferred strategy, while identifying its key assumptions. Only points of information are allowed from other groups at this time. There is then an open debate focusing on which assumptions are different between groups, which are rated differently, and which of the other groups' assumptions each group finds most troubling. Each group should develop a full understanding of the preferred strategies of the others and their key assumptions.

Stage 4 – Synthesis: An attempt to synthesise is then made. Assumptions are negotiated and modifications to key assumptions are made. Agreed assumptions are noted; they can form the basis for consensus around a new strategy that bridges the gap between the old strategies and goes beyond them. If no synthesis can be achieved, points of disagreement are noted and the question of what research might be done to resolve these differences is discussed.

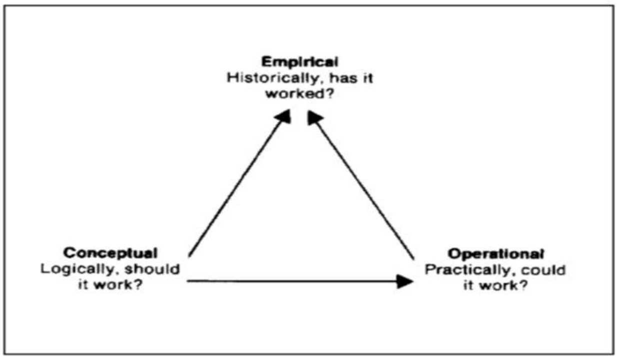
2.6 Prospective Evaluation Synthesis (PES) (GAO, 1995)

In essence, a prospective evaluation synthesis is a combination of: (1) a careful, skilled textual analysis of a proposed programme, designed to clarify the implied goals of the programme and what is assumed to obtain outcomes, (2) a review and synthesis of evaluations from similar programmes, and (3) summary judgments of likely success, given a future context that is not too different from the past. In this respect, the PES resembles the evaluation synthesis approach, except that the focus of the PES is on how evaluation studies cast light on the potential for success of the proposed programmes in the future, as opposed to reaching conclusions about the actual performance of existing programmes.

Conceptually, PES provides a way to use the logic of evaluation methodology and its procedures to assess the potential consequences either of one proposal or of alternative and competing policy proposals. It combines (1) the construction of underlying models of proposed programmes or actions as developed by Wholey for evaluability assessment with (2) the systematic application of existing knowledge as developed in the evaluation synthesis methodology (Wholey, 1977). PES is a prospective analysis anchored in evaluation concepts. It involves operational, conceptual, and empirical analyses, taken in the context of the future (see a figure below).

As the following figure illustrates, the conceptual analyses results help focus the operational analyses and answer the question: 'Logically, should the proposal work?' The operational analyses further scope the search for empirical findings and answer the question: 'Practically, could the proposal work?' The empirical analyses can open both new conceptual and operational possibilities and answer the question: 'Historically, have activities conceptually and operationally similar to the proposal worked in the past?' Finally, the PES takes into account ways in which the past is and is not likely to be similar to plausible future conditions.

Figure 3.1: the Triad of Analysis


Empirical
Historically, has it worked?
Conceptual
Logically, should it work?
Operational
Practically, could it work?

2.7 Elicitation Method

As policies and programmes are developed and implemented by organisations, the 'mental models' or 'cognitive maps' of people in these organisations, i.e., their theories, are important for understanding the anticipated impact of their policies or programmes. The emphasis should therefore be placed on organisational cognitions. One of the central questions is the relationships between these cognitions and the outcomes of organisations. All stakeholders should have 'cognitions' (theories) about the organisation and its environment. These maps of what is going on in their organisation partly determine their behaviour. Their content concerns the organisational strategies, their chances of success, the role power plays, their own roles and the relationships with the outside world. Parts of these maps or theories are implicit and are tacit knowledge, both on an individual and on a collective level. By articulating these mental models, it is possible to compare them with evidence from scientific organisation studies. The articulation is also important for organisations to become 'learners'.

Examples of techniques for reconstructing, eliciting and assessing these mental or cognitive maps are the following:

- Look at the concrete record of strategic intentions, through, for example, a study of the documentation which is designed to direct behaviour;
- Look at decision-making in action; get involved in the organisation (an anthropological observer approach). Watch decision-makers, listen to stories;
- Work with managers on strategic breakdown situations. Become immersed in the thinking and the social process of 'strategic fire fighting';
- Use well-designed trigger questions in interview situations so that 'theories in use' can be detected. Follow interviews with feedback to individuals and to the team. The 'elicitation cycle' is built on responses to designed trigger questions. The process uses six techniques:
    - Create an open-ended atmosphere in the interview;
    - Do away with formal language and create a 'playful' atmosphere in which it is easier to deviate from the formal phraseology and the official script;
    - Do 'set the interviewees up against themselves';
    - Create dialectical tension by asking the interviewees to adopt unusual roles;
    - Listen very carefully for internal inconsistencies in what is being said;

- Apply data/content-analysis programmes or other text analysis programmes to the interview reports and documents; and
- Confront the results of these content-analysis activities with relevant (social) scientific research.

2.8 General Elimination Methodology, also known as Modus Operandi Approach

The core elements of the 'general elimination methodology' are the following (Scriven, 2008):

- The general premise is the deterministic principle: all macro events (or conditions, etc.) have a cause.

- The first 'premise from practice' is the List Of Possible Causes (LOPCs) of events of the type in which we are interested, e.g., learning gains, reduction of poverty, extension of life for AIDS patients. People have used LOPCs for more than a million years, in tracking and cooking and healing and repairing, and today every detective knows the list for murder, just as every competent mechanic knows the list for a brake failure, though the knowledge is as often tacit as explicit, outside the classroom and the maintenance videos. An LOPC usually refers to causes at a certain temporal or spatial remove from the effect, and at a certain level of conceptualisation, and will vary depending on these parameters; of course, the context of the investigation determines the appropriate distance parameters. The distant LOPC for murder is the list of possible motives; a more proximate one, developed in a particular case by applying the general one, is the list of suspects. When dealing with new effects, we may not be certain the list is complete, but we work with the list we have and extend it when necessary.

- The second practical premise is the list of the modus operandi for each of the possible causes (the MOL). Each cause has a set of footprints, a short one if it is a proximate cause, a long one if it is a remote one, but in general the modus operandi is a sequence of intermediate or concurrent events or a set of conditions, or a chain of events, that has to be present when the cause is effective. There is often a rubric for this; for example, in criminal (and most other) investigations into human agency, we use the rubric of means/motives/opportunity to get from the motives to the list of suspects. The list of modus operandi is the magnifying lens that fleshes out the candidate causes from the LOPC so that we can start fitting them to the case or rejecting them, for which we use the next premise.

- The third premise comprises the 'facts of the case,' and these are now assembled selectively, by looking for the presence or absence of factors listed in the modus operandi of each of the LOPCs. Only those causes are (eventually) left standing whose modus operandi are completely present. Ideally, there will be just one of these, but sometimes more than one, which are then co-causes.

3. What and When can Theory-Based Evaluation Contribute?

3.1 Before Implementation

To learn about the plausible effectiveness of a new intervention, an analysis of the theory underlying the intervention can be done. The evaluation tries to open the black box of the intervention: what are the mechanisms that are believed to make the intervention work? How plausible is it that these mechanisms ´do the job´? To detect these mechanisms, one has to search in documents, interviews, transcripts and speeches (of policy-makers, civil servants, etc.) for statements that answer the question why it is believed (or hoped) that the new intervention will make a difference.

It is crucial to be clear about what mechanisms are. Mechanisms are not the input-output-outcome process-variables, nor are they the dimensions usually contained in logical frameworks, logic models or statistical equations. Coleman (1990) and others point to three types of mechanisms: situational, action-formation mechanisms and transformational.

- Situational mechanisms operate at the macro-to-micro level. This type of mechanism shows how specific social situations or events help shape beliefs, desires, and opportunities of individual actors. An example is the opportunity structure a community, village or city - the more there are opportunities (for crime, for unemployed), the larger the chance that crimes will be carried out and jobs will be found.

- Action-formation mechanisms operate at the micro-to-micro level. This type of mechanism looks at how individual behavioral choices and actions are influenced by specific combination of desires, beliefs, and opportunities. Examples are cognitive biases (cognitive dissonance, fundamental attribution error), incentives (rational choice, exchange).

- Transformational mechanisms operate at the micro-to-macro level and show how a number of individuals, through their actions and interactions, generate macro-level outcomes. An example is ‘cascading’ by which people influence one another so much that people ignore their private knowledge and instead rely on the publicly stated judgments of others. The ‘bandwagon phenomenon’— the tendency to do (or believe) things because many other people do (or believe) is related to this, as are ‘group think’ and ‘herd behavior’.

One option to find information on mechanisms is to carefully read documentation and search for statements indicating the (espoused) motivations or rationales behind an intervention ('we believe that…', 'it is generally accepted that this option is …', 'based on our experience with the policy field, we decide that …'; 'the only real option to address this problem is ...'). One can apply content analysis to do this work. However, often mechanisms are not described clearly; they can only be found by reading between the lines and by applying argumentation analysis (Leeuw, 2003). Various argument ('assumption') visualization software applications can be used to detect arguments and order them and (logically) relate them to one another.

Once the mechanisms have been detected, the next step is to compare the statements, assumptions or beliefs (of policy-makers) about mechanisms with evidence from review and synthesis studies. Put differently: compare policy beliefs about mechanisms with research-based evidence. The evidence can be found in repositories like the Campbell Collaboration, the UK Evidence Network, the What Works Clearing House and others (see Hansen & Rieper, 2010 for an overview), but also (meta) search databases like the Web of Science are relevant.

The more the mechanisms believed by policy makers to be at work are in line with research-based evidence, the greater the plausibility of the new intervention to be effective.

Example 1: Subsidies

Pawson (2002) categorized six subsidies, covering incentives to stimulate fire alarm installation in homes, to give up smoking, to widen educational opportunities for students, to improve property, to help ex-offenders to re-socialise and to reduce inner city environmental pollution through subsidizing free-city-centre bikes. Next, he inventoried evaluations of these subsidies and studied the role of situational (context) mechanisms in understanding the success or failure of the subsidies. He produced a list of nine context factors that contribute to success or failure. In order to judge the plausibility that the new subsidy would be effective, Pawson's context factors can be compared to the assumed context mechanisms of the new subsidy. The more the new subsidy takes into account the context mechanisms Pawson found in evaluations of (relatively) successful subsidies, the more plausible it is that the new subsidy will be effective.

Example 2: Fear-Arousal Communication and Behavior Change

This example deals with the fight against cocaine smuggling through people swallowing the drug and travelling between the Dutch Antilles and the Netherlands. Young deprived men were paid through organised crime several thousand Euros to fly between the Antilles and Amsterdam using the internal concealment method, i.e., swallowing small balls filled with cocaine and delivering the 'stuff' in Holland. The Dutch government was successful in reducing this kind of drug trafficking through an almost 100% control of passengers arriving at Amsterdam who came from certain regions. However, the policy was expensive, which made officials to think about an alternative. Could a public information campaign using leaflets, mass media and local media that present fear-arousal information about the medical dangers of the internal concealment method and the likelihood to be arrested, be an effective (and less expensive) intervention to reduce trafficking?

Kruisbergen (2007) evaluated this policy idea. He synthesised results from evaluations of the impact of 'fear-arousal health education programmes' in general (about smoking, dangerous drinking, etc.) and compared the mechanisms and contexts found in these studies with the existing empirical information about the contexts in the Dutch Antilles and some of the social and behavioral characteristics of 'cocaine swallowers'. There was a huge discrepancy between contexts and mechanisms of successful fear-arousal communication health campaigns and the specific characteristics of cocaine swallowers and their contexts. Crucial conditions that made fear-arousal communication have an impact on (health) behavior did not exist in the case of cocaine swallowing behaviour. Kruisbergen's conclusion was that the likelihood of preventing illegal `immigration´ of cocaine to the Netherlands by implementing a public awareness campaign would be small to very small. In other words: the plausibility of the theory that fear-arousal communication will reduce drug trafficking using the internal concealment method was very limited

Example 3: Educational Governance

Janssens & de Wolf (2009) carried out an ex ante evaluation of the theory underlying a new Dutch educational policy that combines accountability and inspections. A central feature of this policy is that it strives for an optimal balance between accountability, inspections, self-evaluation and improvement activities. The programme is called 'educational governance'. It stimulates systems of internal quality assurance by (a) establishing national standards and public accountability, (b) encouraging parents to take an active role in internal supervision processes within schools (through boards of trustees), and (c) enacting external government supervision. One of the objectives is to make schools take a proactive role in educational accountability as opposed to a reactive one. Another objective is to involve other actors (parents, students and teachers) in the accountability system.

The authors applied approaches developed by Pawson and Tilley (1997), Weiss (2000), Leeuw (2003) and others. The evaluation consists of three parts. First, it reconstructs the theory that underlies the aims and policy of educational governance. It identifies the central assumptions of the policy and uses them to reconstruct its causal scheme, 'reconstruction of the programme theory’. The second step is an assessment of the main assumptions of the programme. This involves assessing the tenability of these assumptions in light of the most recent research based insights. With this, the evaluation ascertains the acceptability and empirical tenability of the ideas or assumptions and the validity of the logic underlying the programme theory. The more ‘suitable’ and ‘evidence-based’ the assumptions are, the greater the chance that the programme theory will work in practice. In the last step, the evaluation combines and weighs the conclusions of the evaluations of the separate assumptions. It also determines the mutual compatibility of the assumptions. This last step explores if the programme will be able to generate the intended effects. It also helps to identify theoretical imperfections or other threats to the effectiveness of the programme in practice.

The evaluation found that the policy might not achieve its objectives and identified the elements which needed improvement. A flaw in the theory underlying the programme was found, which threatened its potential effectiveness. Furthermore, the evaluation showed that there was a risk of contrary and incompatible interests among actors, as well as some practical reasons why the programme might not work.

Conclusion

To answer the plausibility question ex ante, it is suggested to focus on mechanisms as the ‘drivers’ of the new policy or intervention and then compare these with already available research and evaluation evidence for these or similar mechanisms. The more the intervention theory is backed by evidence on working mechanisms, the more plausible the theory is and the likelier it is that the new policy will make a difference.

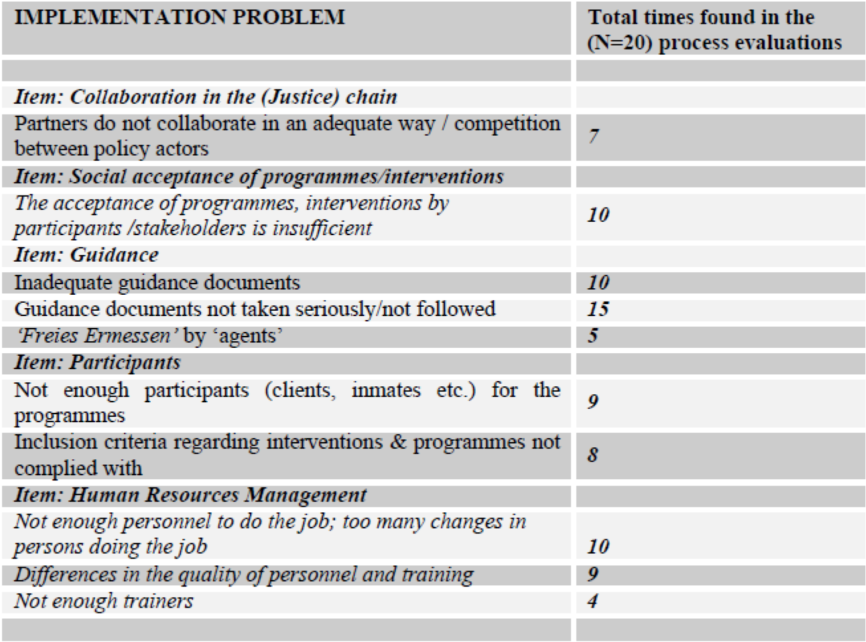
3.2 During Implementation

What can TBE contribute to find out - during implementation - how plausible it is that a policy or programme will be effective? Two routes can be distinguished.

The first route focuses on the implementation theory, i.e., the theory that describes which operations have to be performed and which (organisational) conditions have to be met for a new intervention be ´put to work´. There is abundant evidence that when ‘programme integrity’ is limited, which means that the implementation of the policy is not as was planned, this reduces the effectiveness of the policy (Barnoski, 2004; Carroll, Patterson, Wood, Booth, Rick & Balain, 2007; Nas, van Ooijen & Wieman, 2011). Take the example of a new intervention on e-learning. Such an intervention not only needs to be based on sound (‘working’) mechanisms, but several practical problems also have to be solved. The ICT infrastructure, including bandwidth has to be available and ready; staff, teachers and parents have to accept the new approach; software programs have to be available; students have to work with them and side-effects have to be understood. An example of a side effect is the time it can take to train (older) staff members to become familiar with e-learning and to be able to coach the educational processes.

In a recent Dutch meta-study of 20 implementation evaluations of interventions in the world of crime and justice, the following implementation problems were found to happen most often (Nas, van Ooyen & Wieman, 2011; Leeuw, 2011).

Table 2 Incidence of problems when implementing (penal) sanctions/ behaviour (modification) programmes (based on 20 Dutch process evaluations).



IMPLEMENTATION PROBLEM	Total times found in the (N=20) process evaluations
Item: Collaboration in the (Justice) chain	
Partners do not collaborate in an adequate way / competition between policy actors	*7*
Item: Social acceptance of programmes/interventions	
The acceptance of programmes, interventions by participants /stakeholders is insufficient	***10***
Item: Guidance	
Inadequate guidance documents	***10***
Guidance documents not taken seriously/not followed	***15***
‘*Freies Ermessen’* by ‘agents’	***5***
Item: Participants	
Not enough participants (clients, inmates etc.) for the programmes	***9***
Inclusion criteria regarding interventions & programmes not complied with	***8***
Item: Human Resources Management	
Not enough personnel to do the job; too many changes in persons doing the job	***10***
Differences in the quality of personnel and training	***9***
Not enough trainers	***4***

The implementation of these interventions did not take into account the likelihood that factors such as social acceptance, lack of guidance, collaboration problems and personnel problems would cause problems. The more the theory underlying implementation takes account of these and similar implementation problems, and how to prevent or reduce them, the greater the likelihood of the new programme being successful (if, of course, the intervention theory is plausible). When no information is available on the problems, the plausibility of the intervention being effective is reduced.

The second route for TBE to assist in evaluations during implementation is described by Kautto & Simila (2005) and focuses on the intervention theory and 'recently introduced policy instruments' (RIPI’s). When evaluators are confronted with the request to assess the (future) impact of policies, this is understandably difficult. It takes time for an intervention to be fully implemented and ‘working’. As evaluation time is not similar to political time, this poses problems for policy-makers, evaluation commissioners and evaluators. Kautto and Simila (2005) presented an approach in which a central role is given to the intervention theory. They evaluate a change of the Environmental Protection Act (1999) in Finland. A European Union Directive on Integrated Pollution Prevention and Control was transposed into the Finnish legal system. At the core of the reform was the integration of five different permits (air pollution, water pollution, waste management, protection of health and neighbourhood relations) into one environmental permit. To establish the (final) impact on the environment of the new legal arrangement including the reduction of permits would take years. Waiting that long was not an option, so the evaluators did something else. They started an evaluation relatively soon after the announcement of the new Act. They unpacked the intervention theory (‘why will a reduction from five permits to one be effective in terms of environmental protection?’), distinguished between outputs and outcomes of the Act and collected data on outputs that were already available. Information on outcomes was not yet available. They checked the plausibility of the part of the intervention theory that linked certain characteristics of permits to the final goal of the new Act (environmental impact/outcomes). Because more than 600 permits (= outputs) had been granted during the first two years of implementation, it was possible to assess whether the assumptions about the characteristics of the outputs were correct. Kautto and Simila: 'this enabled us to say something important about the effectiveness despite the fact that the (final) outcomes had not yet occurred. Concurrently, it must be noted that while the permits have not been changed as assumed at the beginning of the implementation process, this does not mean that they will never be changed. The evaluation itself may have an impact on the implementation and as a result, or for other reasons, the authorities may place greater emphasis on gaps and priorities in the future. In this context, the intervention theory was not used to predict the future, but to guide the evaluation'.

An interesting conclusion Kautto and Simila draw is that although an impact analysis was not possible because outcomes had not yet occurred, this does not necessarily imply that the use of impact as a criterion is also impossible – thanks to the concept of an intervention theory. However, the content of the effectiveness criterion must be reformulated. As the evaluators have shown, effectiveness refers to the degree of correspondence between intended policy goals and achieved outcomes. If outcomes have not yet occurred, a comparison of the objectives and achieved outcomes is impossible. But what is possible, instead, is to ask whether the outputs include features that are preconditions to the achievement of the goals according to the intervention theory. The more this is the case, the greater the probability that effectiveness is in reach.

Conclusions

To answer the impact question during the implementation process can be done in two ways. The first is to study the implementation theory and check to what extent this theory takes account of pitfalls that can often be found when implementing interventions. The second route is to follow an approach articulated by Kautto and Simila (2005) known as RIPI-evaluations (‘recently implemented policy instruments’).

3.3 After Implementation

3.3.1 What can TBE contribute ex post to establishing the counterfactual, when it is not feasible to use experimental or non-experimental designs?

If an experimental setting is not possible, if natural experiments or different designs of non-experiments are not possible, then one can move to more qualitative approaches to establish a counterfactual.

- Use the counterfactual history approach and hypothetical question-studies

Counterfactual history, also sometimes referred to as virtual history, attempts to answer 'what if' questions. It seeks to explore history and historical incidents by means of extrapolating a timeline in which certain key historical events did not happen or had an outcome which was different from that which did in fact occur. Fogel (1964) looked at where America would have been (in terms of its GDP) had there been no railroads. He hypothesised that the increase in GDP, given by the railroads, would have happened anyway had other technologies taken hold. Examining transportation costs for primary and secondary goods, he compared the 1890 economy to a hypothetical 1890 economy in which transportation infrastructure was limited to wagons, canals and rivers. Fogel found that the impact of the railroads was small - about 7% of the 1890 GDP. A substitute technology, the more extensive canal system, would have been able to reach a comparable economic growth. After Fogel many other counterfactual historical studies were published, including work that combines experimental psychology and history.

Methodological rules of thumb are available on how to do this work and how to judge its quality (Tetlock & Belkin, 1996). These authors also collect data from hundreds of experts that predict the counterfactual future/past. By analysing their answers, patterns and (ultimately, when time progresses) the validity of their statements, these researchers are trying to unpack what did the 'work' in predicting the future (and vice-versa: the past).

For evaluators, a similar approach is possible. If, for example, the impact of a grant to companies to stimulate innovation or a new system of knowledge brokers for SME has to be assessed and statistical evaluation designs are not possible, evaluators can be asked to develop a counterfactual for the situation had there been no grants. Answering the question can be done in line with the way historians work (using existing data and theories), but it is also possible to apply the hypothetical question-methodology, known from policy acceptance studies and marketing. The question then is what people would do if policy \`a\` or \`b´, was not implemented. An early example is the Thompson & Appelbaum (1974) study of the impact of population policies in the USA, which was later one of the pillars upon which Dutch hypothetical question evaluations of population policies were built (Moors et al, 1985).

- Apply contribution analysis

Contribution analysis is based on the existence of, or more usually, the development of a theory of change for the intervention being examined. A theory of change sets out why it is believed that the intervention’s activities will lead to a contribution to the intended outcomes; that is, why and how the observed outcomes can be attributed to the intervention. The analysis tests this theory against logic and the evidence available on the outcomes observed and the various assumptions behind the theory of change, and examines other influencing factors. It either confirms the postulated theory of change or suggests revisions in the theory where the reality appears otherwise. It is best done iteratively, building up over time a more robust contribution story. The overall aim is to reduce uncertainty about the contribution the intervention makes to the observed outcomes through an increased understanding of why the outcomes have occurred (or not occurred) and the roles played by the intervention and other factors. Mayne (in press) has outlined steps of contribution analysis including examples in different fields (see section 2.3).

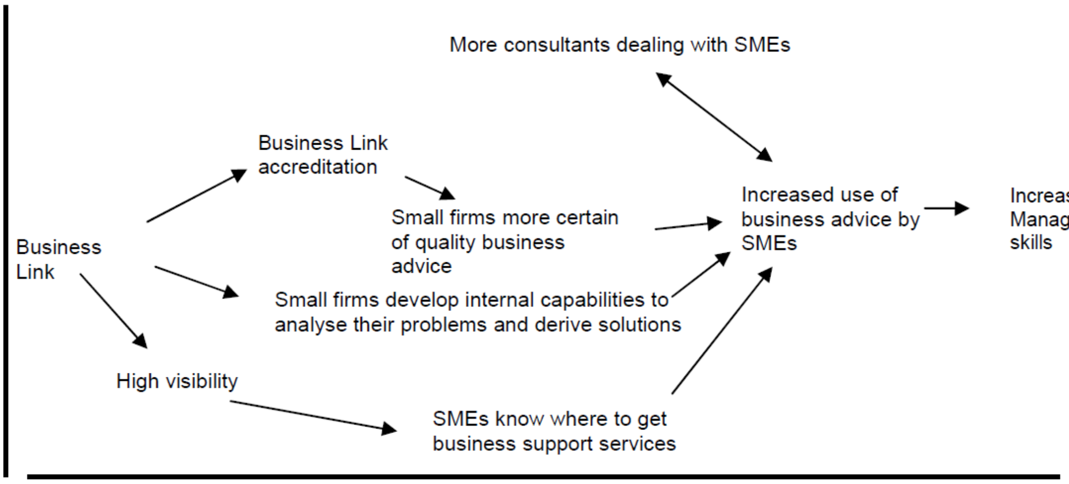
Although the authors did not relate their study to contribution analysis, Mole’s et al (2007) large-scale telephone survey of over 3,000 SMEs and 40 face-to-face interviews with business owners, which tried to assess the impact of Business Link Organization (BLO) activities on businesses that received assistance, is somewhat linked to this approach.

Evaluating the Impact of Businesses Link Organization activities (Mole et al, 2007)

BLO is a type of small business support activities that some (European) governments have implemented. It can be seen as a type of brokerage and is usually targeted at small (to medium size) businesses. BLO activities, in the end, are believed to increase economic productivity and job growth. The small business service aims to build the capability for small business growth and the advice and support provided by Business Links are intended to improve the management skills and thereby improve business performance and entrepreneurship.

The evaluation by Mole et al (2007) paid some attention to the programme theory underlying the BLO (see figure below). In the figure the arrows indicate a direction of causality (Mole et al, 2007: 28).

Figure 1.1: Programme Theory for Business Links


More consultants dealing with SMEs
Business Link accreditation
Small firms more certain of quality business advice
Increased use of business advice by SMEs
Increased Management skills
Business Link
Small firms develop internal capabilities to analyse their problems and derive solutions
High visibility
SMEs know where to get business support services

Although the evaluators refer to this programme theory and also present several hypotheses on relationships between BLO and dependent variables like increased management skills and the possibility of finding business advice, no attention was paid to the question how and to what extent the programme (theory) is empirically linked to policy goals like productivity increase or job creation. Had these topics been added to the (large scale) empirical approach of the evaluation, it would have made the relevance of the study larger[2].

- Work in line with expert judgments

Expert judgments or connoisseur evaluations are used to cover strategies that pool the opinions of experts to assess performance of programmes or policies. Recent forms of expert judgment include:

[2] Also, a number of other (methodological) problems in the study would have to be addressed, like the quality of data (collection).

- Accreditation and evaluation activities of the effectiveness of behavioural modification programmes or educational programmes. This is a combination of expert judgment and meta-evaluation & meta-analysis of what is known about – for example - programmes reducing violence in public places by people.
- Civic evaluation, based on 'the wisdom of the crowd' that evaluates policies and organisations. Here the experts are 'the people' those engaged in these social groups.

If one wants to use this approach to develop the counterfactual, it must be stressed that the evaluators coordinating this effort are not looking for answers from experts on the question what the impact of intervention X would have been, but exactly on the opposite question: what would have happened without intervention X.

- Work in line with the GEM: 'general elimination methodology' when using one of these approaches.

GEM was coined by Scriven (1978; 2008) – see section 2.7. If an evaluation has found results on impact although the design is weak or insufficient, and if the underlying intervention theory is relatively plausible, the GEM can be used to check if there are other factors (than the intervention) that are more plausible as explanations of the impact. The primary goal of GEM is to see how solid the arguments are, indicating that the intervention caused or contributed to the outcomes. A GEM evaluator invites the 'believers' in the contribution of the intervention to discuss alternative explanations, having nothing to do with the intervention. The more believers serious challenge and falsify these alternative explanations, the more plausible it is that the intervention indeed is causing the difference. Simultaneously the GEM evaluator tries to falsify the intervention theory. The more successful they are in doing that, the less plausible this theory is.

3.3.2 What can TBE contribute ex post when an impact evaluation, including the counterfactual established through experimental or non-experimental designs, has been carried out, but an explanation of the findings is lacking?

Evaluators applying experimental or non-experimental designs do not always pay attention to the social and behavioural mechanisms that underlie the interventions they assess. The interventions, Pawson and Tilley (1994) claim, are seen almost as black boxes, whereas to understand why things work (assuming they do), one needs to know which social and behavioural mechanisms are active and in which contexts (Pawson & Tilley, 1997). During the 1990s and early 2000s an almost paradigmatic conflict existed between (some) experimentalists and (some) realists on this topic. More recently this conflict has become less severe.

What can be done to remedy the lack of explanations? The first answer is to open up the black boxes afterwards in the way suggested in section 3.1. The second answer is to combine experimental and non-experimental impact evaluations and realist evaluations. An example of this approach can be found in a paper by Van der Knaap et al (2008). It describes an approach that combines the use of the 'Campbell collaboration standards' with the realist notion of addressing contexts-mechanisms-outcomes (CMO) that underlie interventions (see section 2.1 on realist evaluation).

The Campbell Collaboration (C2) is an international volunteer network of policymakers, researchers, practitioners, and consumers who prepare, maintain and disseminate systematic reviews of studies of interventions in the social and behavioral sciences (see http://www.campbellcollaboration.org). The organisation is named after Donald T. Campbell, an American social scientist and champion of public and professional decision-making based on sound evidence. C2 reviews are designed to generate high-quality evidence in the interest of providing useful information to policy-makers, practitioners and the public on what interventions help, harm or have no detectable effect. The organisation has developed standards for systematic review (clear inclusion/exclusion criteria, an explicit search strategy, systematic coding and analysis of included studies, meta-analysis).

Van der Knaap et al focused on interventions to prevent or reduce violence in the public domain. To merge 'Campbell standards' and the realist evaluation approach, the realist approach was applied after finishing the Campbell-style systematic review. The following box describes the way the ‘merger’ works:

'Our first goal was to provide an international overview of effective or at least promising measures to prevent violence in the public and semi-public domains. The second goal was to gain insights into the behavioural and social mechanisms that underlie effective or promising prevention measures and the circumstances in which these are found to be effective. We defined violence as “the deliberate use of physical strength or power and/or the threat thereof, aimed against another person or group of persons and which results or is likely to result in injury, death or psychological damage"' (van der Knaap et al., 2006, p. 21).

'Our first step was to conduct a Campbell-style review. …….We collected 48 studies that met our inclusion and exclusion criteria. These 48 publications relate to 36 interventions, most of which are designed to prevent violence in schools. We did not include a meta-analysis in our study but instead assessed each study’s methodological quality using the Maryland Scientific Methods Scale (MSMS). Only experimental (level 5), quasi-experimental (level 4), comparative designs without matching or randomisation (Level 3) and evaluations using a before-after design (level 2) were used. 13 Level 2 studies, 10 Level 3 studies, 13 Level 4 studies, and 11 Level 5 studies therefore were included. For information on the MSMS, see http://www.ncjrs.gov/pdffiles/171676.PDF).

Based on the MSMS scores, we classified each of the 36 interventions into one of the categories of effective, potentially effective, potentially ineffective, and ineffective. However, not all studies could be grouped into one of the four categories. In 16 cases, the quality of the study design was not good enough to decide on the effectiveness of a measure. Nine interventions were labelled effective and six were labelled potentially effective. Four interventions were labelled potentially ineffective and one was labelled ineffective in preventing violence.

After finishing our Campbell-style review, we applied the realist approach to each of the interventions in our study. This proved to be rather difficult, for a lot of information was missing in the original publications. Often, no explicit theory was described underpinning the intervention, and information on mechanisms and context was scarce. By having two researchers read the publications and identify implicit and explicit statements pertaining to mechanisms and context, we tried to reconstruct CMO configurations. Among other strategies, we scrutinized the outcome measures that were used by the evaluators. For instance, if they focused on attitudes and knowledge, we argued that the program designers meant to achieve changes in attitudes and knowledge and assumed that these changes would cause behavioural change. Whereas some publications offered more detailed information, the mechanisms we identified could mostly be described in general terms only. Based on the evaluations we analysed, some ten mechanisms could be identified that, in fact, boiled down to the following three:

The first of these is of a cognitive nature, focusing on learning, teaching, and training. The second overarching mechanism concerns the way in which the (social) environment is rewarding or punishing behaviour (through bonding, community development, and the targeting of police activities). The third mechanism - of a more general nature - is risk reduction, for instance, by promoting protective factors' (Van der Knaap et al, 2008:55).

Van der Knaap et al (2008) summarise their view on the practical importance of their work as follows: 'Combining the approach outlined by the Campbell Collaboration and the realist evaluation approach is commendable in several ways. First, the result of applying Campbell standards helps to distinguish different types and (methodological) levels of evaluation designs. For those interested in the impact or effectiveness of interventions, this is important. Second, the opening up of the micro-architecture of those interventions that have been shown to be effective, or at least potentially effective, helps better understand what makes these interventions work. Moreover, by also studying the mechanism-context configurations of interventions that appear to be ineffective, one can learn more about the conditions that are necessary for mechanisms to work. A third advantage of our combination, which we have ourselves not been able to realize, is that by applying a realist synthesis approach (Pawson, 2006), knowledge from outside the field of crime and justice evaluations but of direct relevance to the mechanisms can be used to understand why (some) programs work and others do not. In the longer run, these combinations of knowledge funds will help in understanding the interventions better and will probably also help in designing better interventions'.

Conclusions

When an experimental or non-experimental counterfactual is not feasible, TBE can help to derive a counterfactual in several ways. One is to apply the approach of counterfactual historians; another to use hypothetical-question studies while a third is to involve expert judgements ('connoisseur evaluations'). The General Elimination Methodology is also recommended.

When an experimental or non-experimental impact study has been done and results on the effectiveness of the policy or programme are available, attention is not always paid to the why and the how question. These questions are important for policy makers. TBE can help to find explanations: first, by opening up the black boxes of the interventions evaluated and searching for working mechanisms; second, by doing a (follow-up) study in which evaluation designs working with experimental or non-experimental counterfactuals are combined with the realist evaluation approach from the start.

3.4 Indicators

What can TBE contribute to define and operationalise performance indicators of policies or programmes?

TBE and performance indicators are related, but it is a rather complex relationship. There are at least three sets of theories involved.

The first is the theory that the improvement of performance of organisations is stimulated by working with indicators. Mechanisms are that indicators drive workers and management to work (more) in line with the goals set by the organisation and in such a way that comparisons between divisions, departments and outside organisations are possible. It is also believed that indicators stimulate 'learning'.

A second theory specifies the form, content and number of indicators and points to intended and unintended effects of working with them. Some indicators may be better in contributing to learning, while others may stimulate bureaucratization, red tape, dramaturgical compliance and the performance paradox (van Thiel & Leeuw, 2002).

We do not deal with these two sets of theory here. Instead we discuss the question of what policy/programme theories can contribute to developing and implementing effective indicators.

As working with indicators has become mainstream in policy fields and as evidence is available on 'pathologies' that go hand in hand with using indicators (Bouckaert & Balk 1991; van Thiel & Leeuw, 2002), the role policy/programme theory plays in the world of indicators is important.

Fifteen years ago Bickman (1996) suggested that a logical starting point for developing the most appropriate indicators was to create a model or programme theory. Birleson, Brann & Smith (2009) did exactly that in a paper on clinical and community care in Child and Adolescent Mental Health Services (CAMHS) in hospitals. They articulated the programme theory of different services by looking at programme operations, proximal outcomes and final outcomes and relationships between them. They showed that without an articulated programme theory, indicators were likely to be less relevant, over inclusive or poorly linked with the programme operations they aimed to measure. In a rather different field (road safety performance) Hakkert, Gitelman & Vis (2007) did something similar. This study provides details about the theory behind the development of Safety Performance Indicators in seven major areas which are central to the fields of activity in road safety in Europe.

A third example is different as it shows how problematic the use of indicators can be if there is discrepancy between the programme theory and the indicators. Lindgren (2001) shows how thin the ice is for performance measurement, when indicators are developed and used without taking notice of the (richness of the) programme theory. The case concerns popular adult education in Sweden and demonstrates that important activities and characteristics of adult education are not covered by the key indicators, which leads to pitfalls in the use of the performance data, while there is also a set of indicators not linked to any substantive part of the programme theory.

In linking TBE with performance indicators, four activities are central:

- The first is to (re)construct the theory underlying the policy or programme and to develop indicators that cover the richness of the underlying theory.
- The second activity is to understand that indicators can trigger behavioural responses that can lead to a ‘performance paradox’: organisations good in measuring performance indicators are not necessarily the most effective organisations. Examples of these trigger mechanisms are the following:
    - Emphasising - by policy-makers/principal - that compliance with protocols and procedures is crucial (often leading to the production of data largely to satisfy the principals’ need for sound protocols and procedures);
    - Having to work with elusive and contradictory policy goals;
    - Having to work with goals that are inherently not or very difficult to operationalise and measure.
- These trigger mechanisms can contribute to an unintended performance paradox. Van Thiel & Leeuw (2002) also point to the problem that there are mechanisms leading to an intended performance paradox. These are 'cognitive sabotage' of performance measurements and audits, including 'cooking' the data, 'creaming' (focussing on the best cases) and myopia (only information on short term objectives is presented, while more information is available).
- The final step is to prevent the performance paradox. Meyer and Gupta (1994) recommend the use of targets and comparisons over time, between organisations or between different units within the same organisation.

Conclusions

The more important performance indicators are, and the more there is evidence that working with them can have unintended and undesirable side effects, the more it is relevant that TBE is used when designing and implementing them. If not, the likelihood that indicators are distanced from the operations and mechanisms of the policies and programmes analysed will increase.

These conclusions apply directly to EU Cohesion policy programmes. With a growing demand for a more performance-oriented EU Cohesion policy, the importance of performance indicators also increases. This requires a greater focus to be put on indicators which should reflect the objectives of the policy and better capture the effects of the interventions. This new approach is promoted in the context of the 2014-2020 programming exercise (for more information, please refer to:

http://ec.europa.eu/regional_policy/impact/evaluation/performance_en.cfm).

4. Problems to be Avoided when doing TBE?

The following pitfalls or problems when doing TBE can be mentioned. If evaluators are not aware of them, TBE will create 'error costs'[3].

- Avoid sloppy reconstructions and tests of underlying programme theories.

Tilley (1992) brought together several practices that contribute to the production of error costs when doing evaluations Sloppy reconstructions and tests of underlying programme theories (`misconstrue programmes') is one; neglecting contextual differences when comparing results from evaluations (in different time periods) is another. Misinterpretation of what caused a programme not to work (by confusing implementation problems, measurement problems and difficulties with the programme theory) is another (Tilley, 1992). Programmes that could have been effective are sometimes terminated or considered not ready for implementation because of a faulty theory-reconstruction. Error costs involved are inefficiency, foregone investments in developing the programme and wasted money on behalf of the evaluation, while an opportunity cost is that the social problem to be remedied by the programme, continues to exist. Related to this is what Funnel and Rogers (2011) call the 'No Actual Theory' trap: an evaluator refers to programme theories which are in fact not theories at all. '[Instead], they simply display boxes of activities and boxes of outcomes without demonstrating logical and defensible relationships between them and the various items listed in the boxes'.

- Take notice of the problem of concatenation of mechanisms and try to solve it.

Hedstrom (2005) has dealt with this point: 'it is often necessary [when doing a TBE] to consider several mechanisms simultaneously in order to make sense of a specific social phenomenon'. He adds that 'these mechanisms may interact with another in a complex way' (ibid). In recent work by Rogers (2009) and Rogers & Funnel (2011), attention is paid to the relationship between complexity, complicatedness and theory-based evaluations.

- Prevent 'designed blindness'.

This happens when practitioners and evaluators are focused on the programme theory in such an intense way that they not only start to frame every activity of the evaluated programme in

[3] See Leeuw (2010) in the 'Zeitschrift für Evaluation' on costs and benefits of evaluations.

terms of this theory (Friedman, 2001), but also start to believe that the intervention theory is inherently ‘valid’ and ‘good’; this point is related to the psychological mechanisms of tunnel vision. The error cost is that the evaluation ends up being circular: as the evaluator and the evaluated programme are ‘captured´ in the programme theory, the possibility for a serious test of the theory by collecting data for example, is very small.

- Prevent the ‘polishing’ up or quasi-enrichment of the policy theory.

This happens when policy-makers ask evaluators to polish up or ‘enrich’ assumptions underlying their policies, while in reality the policies are grounded on rather thin assumptions. The error costs are twofold: first, it resembles impression management (the 'rich' and informative intervention theory forms the fundament of an intervention that is largely a ‘show policy’) and, secondly, it can set in motion a process of imitation in organisations that will create future failures and faulty processes.

- Not using the programme theory for evaluation.

Funnel and Rogers (2011) refer to this ‘trap’. It concerns the discrepancy between developing or reconstructing the programme theory but nevertheless doing the (empirical) evaluation without paying attention to this theory. It can be labelled a case of ‘wasted words’.

What Theory Based Evaluation is not?
TBE is not the same as presenting:

- A logical framework;
- 'Unexplained causal arrows' and
- Schemes such as the input-throughput-output-diagram often used in (performance) measurements and auditing (Astbury & Leeuw, 2010).

5. Bibliography

Astbury, Brat & Frans L Leeuw (2010). Unpacking black boxes: mechanisms and theory-building in evaluation. American Journal of Evaluation 31 (3): 363-381.

Barnoski, R. (2004). Outcome evaluation of Washington State's research-based programs for juvenile offenders. Washington State Institute for Public Policy.

Birleson, P., Brann, P. & Smith, A. (2001). Using program theory to develop key performance indicators for child and adolescent mental health services. Australian Health Review, 24 (1): 10-21.

Bouckaert, G.& Balk,W. (1991). Public productivity measurement: Diseases and cures. Public Productivity & Management Review, 15 (2): 229-235.

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. Implementation Science, 2(40). http://implementationscience.com/content/pdf/1748-5908-2-40.pdf

Carvalho, S. & White, H. (2004). Theory-based evaluation: the case of social funds. American Journal of Evaluation, 25, (2): 141-160.

Chen, H. T., & Rossi, P. H. (1980). The multi-goal, theory-driven approach to evaluation: A model linking basic and applied social science. Social Forces, 59, 106-122.

Coleman, James (1990). Foundations of Social Theory, Belknap Press.

Donaldson, S. I. (2007). Program theory-driven evaluation science. New York, NY: Lawrence.

Fogel, R. (1964). Railroads and American Economic Growth: Essays in Econometric History (1964).

Hakkert, A.S, Gitelman, V. and Vis, M.A. (Eds.) (2007). Road Safety Performance Indicators: Theory. Deliverable D3.6 of the EU FP6 project SafetyNet.

Friedman, Victor, Designed Blindness: An Action Science Perspective on Program Theory Evaluation. American Journal of Evaluation, 22:161–181.

Hansen, H. & O. Rieper (2010). Institutionalization of Second-Order Evidence-Producing Organizations, in O. Rieper et al. (eds), The Evidence Book: Concepts, Generation and the Use of Evidence, pp. 27-52.

Janssens, Frans & Inge de Wolf (2009).Analysing the Assumptions of a Policy Program. An Ex-ante Evaluation of ‘‘Educational Governance’’ in the Netherlands. American Journal of Evaluation 30(3): 411-425.

Kautto, P. and Similä, J. (2005). Recently Introduced Policy Instruments and Intervention Theories. Evaluation, 11(1): 55–68.

Kruisbergen, E.W. (2005). Voorlichting: doen of laten? Theorie van afschrikwekkende voorlichtingscampagnes toegepast op de casus van bolletjesslikkers. Beleidswetenschap, 19, 3.

Leeuw, Frans L. Policy theories, knowledge utilization, and evaluation. Knowledge and Policy, 4: 73-92.

Leeuw, Frans L. & J.E. Furubo (2008). Evaluation systems: what are they and why study them? Evaluation, 14 (1): 157-169.

Leeuw, Frans, (2009). Evaluation Policy in the Netherlands. New Directions for Evaluation, 123: 87-103.

Leeuw, Frans L. (2003). Reconstructing program theories: methods available and problems to be solved. American Journal of Evaluation, 24 (1): 5-20.

Leeuw, Frans L. (2011). Can legal research benefit from evaluation studies? Utrecht Law Review, 7 (1): 52-65.

Leeuw, Frans & J. Vaessen (2009). Impact evaluation and development. NONIE & World Bank, Washington.

Lindgren, L. (2001). The Non-profit Sector Meets the Performance-management Movement; A Programme-theory Approach. Evaluation, 7(3): 285–303.
Ludwig, J. Kling, J.R, and Mullainathan, S, (2011). Mechanism Experiments and Policy Evaluations. Journal of Economic Perspectives, 25 (3): 17–38.

Mole, Kevin et al, Economic Impact Study of Business Link Local Service, University of Warwick, 2007.

Nas, Coralijn, Marianne M.J. van Ooyen-Houben & Jenske Wieman (2011). Interventies in uitvoering. Wat er mis kan gaan bij de uitvoering van justitiële (gedrags)interventies en hoe dat komt. WODC Memorandum, Den Haag.

Patton, Michael (2008). Advocacy Impact Evaluation. Journal of multidisciplinary Evaluation, 5 (9): 1-10.

Pawson, Ray (2002), Evidence-based Policy: The Promise of `Realist Synthesis´. Evaluation 8 (3): 340–358

Pawson, Ray (2003). Nothing as practical as a good theory. Evaluation 9(4): 471 – 90.

Pawson, Ray (2006 a). Evidence-based policy: a realist perspective. London.

Pawson, Ray (2006 b). Simple principles for the evaluation of complex programmes,' in A Killoran and A Kelly (eds), Evidence based public health . Oxford: Oxford University Press.

Pawson, Ray & Nick Tilley (1997). Realistic Evaluation, London.

Pawson, Ray & S. Sridharan (2010). Theory-driven evaluation of public health programmes, in: Evidence-based Public Health Effectiveness and efficiency, Edited by Amanda Killoran and Mike Kelly, chapter 4: 42-62.

Rogers, Patricia and Sue Funnell (2011). Purposeful Program Theory: Effective Use of Theories of Change and Logic Models. Jossey Bass.

Rogers, P. J. (2008). Using programme theory to evaluate complicated and complex aspects of interventions. Evaluation, 14, 29-48.

Rozendal, P., H. Moors & F. Leeuw (1985). Het bevolkingsvraagstuk in de jaren 80; opvattingen over overheidsbeleid, Nidi, Den Haag.

Scriven, M. (2008). Summative Evaluation of RCT Methodology: An Alternative Approach to Causal Research. Journal of Multidisciplinary Evaluation 5(9), 11–24.

Scriven, M. (1976). Maximizing the Power of Causal Investigations: The Modus Operandi Method, in: G. V. Glass (ed.) Evaluation Studies Review Annual, Vol. 1, Sage Publications, Beverly Hills, CA.

Suchman, E. (1967). Evaluative research. New York, NY: Russell Sage Foundation.

Tetlock, Philip E. and Aaron Belkin (1996). “Counterfactual thought Experiments in Global Politics: Logical, Methodological, and Psychological Perspectives.” In Tetlock and Belkin (eds) Counterfactual Reasoning, Counterfactual thought experiments in global politics: Logical, Methodological and Psychological Perspectives. Princeton University Press, pp. 3-38.

Thiel, Sandra van & Leeuw, Frans L. (2002). The performance paradox in the public sector. Public Productivity and Management Review, 25: 267-281.

Thompson, V.D. & Appelbaum, M. (1974). Population Policy Acceptance: Psychological Determinants. Chapel Hill, N.C.: Carolina Population Center Monograph Series.

Tilley, Nick (1999). Evaluation and evidence-(mis)led policy. Evaluation Journal of Australasia, 11: 48-63.

US GAO (1995). Prospective evaluations methods, Washington.

US GAO (1986). Teenage pregnancy. 500,000 Births a year but Few Tested Programs. Washington.

Van der Knaap, Leontien M., Frans L. Leeuw, Stefan Bogaerts and Laura T. J. Nijssen (2008), Combining Campbell Standards and the Realist Evaluation Approach: The Best of two world. American Journal of Evaluation, 29 (1): 48-57.

Weiss, C. H. (1995). Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. In J. Connell, A. Kubisch, L. B. Schorr, & C. H. Weiss (Eds.), New approaches to evaluating community initiatives: Volume 1, concepts, methods, and contexts (pp. 65-92). New York, NY: Aspen Institute.

Weiss, C. H. (2000). Which links in which theories shall we evaluate? In P. J. Rogers, T. A. Hasci, A. Petrosino, & T. A. Huebner (Eds.), Program theory in evaluation: Challenges and opportunities (pp. 35-45). New Directions for Evaluation, No. 87. San Francisco, CA: Jossey-Bass.